Saliency prediction via multi-level features and deep supervision for children with autism spectrum disorder

Weijie Wei, Zhi Liu, Lijin Huang, Alexis Nebout, Olivier Le Meur

SALIENCY PREDICTION VIA MULTI-LEVEL FEATURES AND DEEP SUPERVISION FOR CHILDREN WITH AUTISM SPECTRUM DISORDER

Weijie Wei[†‡], *Zhi Liu*[†‡*], *Lijin Huang*[†‡], *Alexis Nebout*[§], *and Olivier Le Meur*[§]

[†]Shanghai Institute for Advanced Communication and Data Science, Shanghai University, China
[‡]School of Communication and Information Engineering, Shanghai University, China
[§]IRISA, University of Rennes 1, Rennes 35042, France
codename1995@shu.edu.cn, liuzhisjtu@163.com, hyrx@live.cn, alexis.nebout@irisa.fr, olemeur@irisa.fr

ABSTRACT

This paper proposes a novel saliency prediction model for children with autism spectrum disorder (ASD). Based on the convolutional neural network, the multi-level features are extracted and integrated to three attention maps, which are used to generate the predicted saliency map. The deep supervision on the attention maps is exploited to build connections between ground truths and the deep layers in the neural network during training. Furthermore, by performing the single-side clipping operation on the ground truths, our model is encouraged to enhance the capacity of better predicting the most salient regions in images. Experimental results on an ASD eye-tracking dataset demonstrate that our model achieves the better saliency prediction performance for children with ASD.

Index Terms—Saliency prediction, visual attention, saliency model, autism spectrum disorder.

1. INTRODUCTION

People with ASD perform atypically with respect to the controls while they are viewing the world. People with ASD will pay more attention to idiosyncratic objects and less focus on social objects (*e.g.* face) than normal humans do [1]. In [2], a three-layered saliency model is proposed to quantitatively measure the difference between ASD group and controls, and the results show that people with ASD have a stronger center-bias and less attention on faces and socially gazed regions. The study in [3, 4] show that such abnormal visual attention can be found even in children and adolescents. By collecting eye-tracking data from children with ASD while they are watching videos delivering actual courses, it is found that the unwillingness to gaze at teachers and the persistence in other areas lead to their maladaptation to school [5]. Studying this kind of atypical visual attention would help design textbooks or CHIs (Computer Human Interfaces) for people with ASD.

The distinctive characteristic of visual attention has been regarded as a valid biomarker to assist clinicians for better ASD diagnosis [6]. In [7], children's eye-tracking data while gazing at several short videos were exploited to discriminate ASD from TD (typical development). In [8], the difference of fixation density maps, which were generated by eye-tracking data, between ASD group and TD group, was exploited to drive the model generating the most discriminative features for an effective classification.

The performance of saliency prediction makes a huge progress due to the breakthrough of deep learning technique and ever-growing datasets. SALICON [9] and ML-Net [10] re-purpose the existing deep architectures by the convolutional neural network (CNN) with two-path shared weights and multi-level features, respectively. A skip-layer network, which use hierarchical saliency information to refine coarse and local saliency response, is proposed to predict pixel saliency [11]. In [12], the attentive convolutional long short-term memory (ConvLSTM) is integrated into a saliency attentive model for fixation prediction. The state-of-the-art saliency prediction models mentioned above can generate predictions close to human fixation maps [13].

In this paper, we focus on predicting visual attention of children with ASD. Specifically, our contributions can be summarized in twofold:

(1) We propose a novel saliency prediction model that exploits multi-level features and deep supervision to effectively predict the fixations of children with ASD.
(2) To facilitate the proposed model to focus on learning the features of salient regions, the fixation density maps as the ground truths are processed via single-side clipping, which further boosts the saliency prediction performance.

The rest of this paper is organized as follows. The proposed saliency prediction model is detailed in Section 2. Experimental results and analysis are presented in Section 3, and conclusions are drawn in Section 4.

This work was supported by the National Natural Science Foundation of China under Grant No. 61771301.

[*]Corresponding author: Zhi Liu.

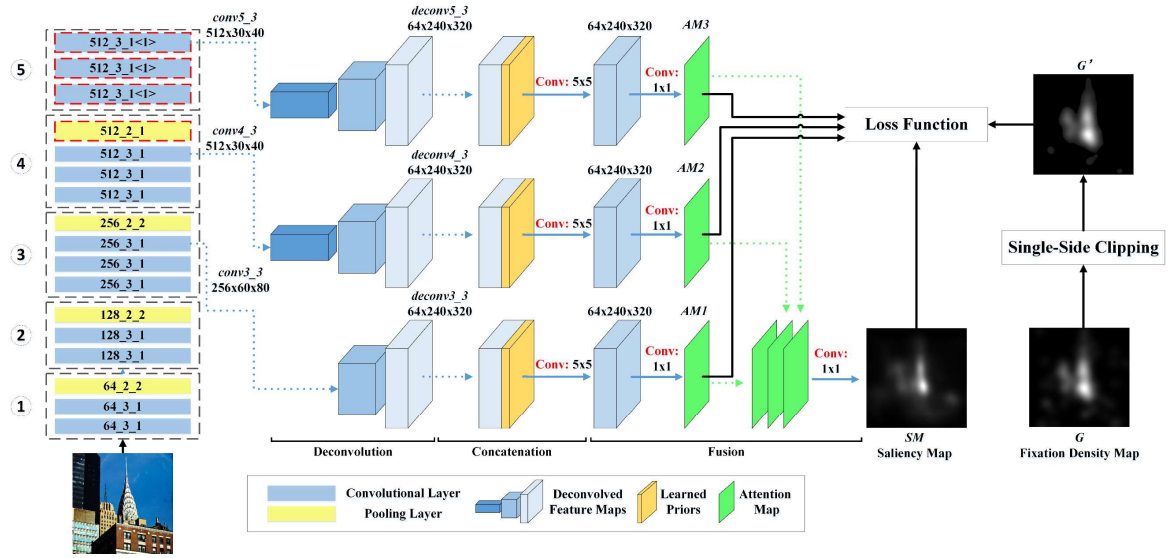

Fig. 1. Overview of the proposed network architecture. The left part is the dilated convolutional network (DCN) where the layers are expressed in terms of *channels_kernel_stride<holes>*. The red dashed boxes indicate the modified layers.

2. SALIENCY PREDICTION MODEL

The CNN is proved to be an effective architecture to predict human fixations. CNN is capable of capturing both low-level contrast features and high-level semantic features, which describe the characteristics of the original image and help generate the predicted saliency map.

As shown in Fig. 1, we fully adopt the multi-level features derived from the dilated convolutional network (DCN) and utilize deconvolutional operations to upsample feature maps through multiple paths. The prior maps learned from ground truths are combined with the upsampled maps by using convolutional operations. The outputs of this operation are three attention maps, which are concatenated and convolved to generate the saliency map, as the output of our model. In the training phase, both attention maps and saliency map are used to calculate the loss, which is exploited to optimize our model with deep supervision manner.

2.1. Dilated Convolutional Network

VGG-16, which contains 13 convolutional layers and 3 fully connected layers, is an outstanding architecture for image classification [14]. Since the saliency prediction is an image-to-image task, the last three fully connected layers of VGG-16 are removed to avoid the condensation of spatial information. Besides, the pooling layers that downsample image will worsen the performance. To mitigate the influence of reduced resolution, the last max pooling layer is removed and the stride of the penultimate max pooling layer is set to 1. Besides, the kernels of three convolutional layers in the 5th block are replaced by the dilated kernels [22] whose holes are set to 1, as shown in the red dashed boxes of Fig. 1. The outputs of the last convolutional layers in the 3rd, 4th and 5th block are denoted as *conv3_3, conv4_4 and conv5_3*, respectively. These feature maps are fed into the deconvolution layers in multiple paths as shown by the dotted lines in Fig. 1.

2.2. Deep Supervision

We exploit the deconvolution layers to upsample feature maps, in order to supervise not only the output of the last convolutional layer but also the deep convolutional layers in the network. The deconvolution operation on feature maps, which reduces the number of channels by a factor of two while increases the spatial resolution by a factor of two, can be regarded as the reverse operation of convolution. The outputs of the three deconvolution layers are denoted as *deconv3_3, deconv4_3* and *deconv5_3*, respectively. All the three deconvolved feature maps share the same dimension as shown in Fig. 1.

The learned prior maps come from the self-learning block in [12], which learns the parameters of 2D Gaussian distribution to fit the location-bias of the dataset. *Deconv3_3, deconv4_3* and *deconv5_3* are concatenated with the learned prior maps. The resulting three tensors pass through the convolutional layers with 64 filters and 1 filter in succession to generate the three attention maps, *i.e. AM1, AM2* and *AM3*, respectively. Finally, the predicted saliency map *SM* is generated by convolving the concatenated attention maps. The loss function is defined as follows:

$$L(P, G', FP) = \sum_i [F_1(P_i, G') + F_2(P_i, G') + F_3(P_i, FP)] \quad (1)$$

where P_i indexes four predictions, namely *SM*, *AM1*, *AM2* and *AM3*, *G'* and *FP* are the clipped ground truth and fixation points map. F_1, F_2 and F_3 are respectively the Pearson correlation coefficient (CC), the Kullback-Leibler divergence (KL) and the normalized scanpath saliency (NSS). Specifically, they are calculated as follows:

$$F_1(P_i, G') = -2 \times \frac{\sigma(P_i, G')}{\sigma(P_i) \cdot \sigma(G')} \quad (2)$$

where $\sigma(\cdot)$ and $\sigma(\cdot,\cdot)$ represent standard deviation and covariance, respectively. Since CC has a higher value when two variables are more similar, it should multiple a negative coefficient while being added to the total loss.

$$F_2(P_i, G') = -\sum_j^N G'_j \log\left(\frac{G'_j}{P_{i,j}}\right) \quad (3)$$

where j indicates j^{th} pixel and N is the total number of pixels. For the same reason, F_2 should multiple a negative coefficient.

$$F_3(P_i, FP) = 10 \times \frac{1}{N} \sum_{k=1}^{M} \frac{P_{i,k} - \mu(P_i)}{\sigma(P_i)} FP_k \quad (4)$$

where M is the total number of fixations. The map P_i is normalized to zero means and unit standard deviation. The above defined loss function encourages the ground truth to supervise deeper features in the network and to train a high-performing saliency prediction model.

2.3. Single-Side Clipping

Generally, the human fixation map is a binary map of fixation location. The fixation density map (FDM), as the ground truth to train the saliency prediction model, is generated by smoothing the fixation map with a Gaussian filter and performing a min-max normalization. Thus, the most salient region, which is gazed by most observers, has dense fixations. For the task of saliency prediction for children with ASD, what we concern is the attractive regions for observers with ASD. There are two sets of such attractive regions. The first set includes regions that attract attention from both ASD and TD. The features of this kind of regions can be learned through the training on large-scale eye-tracking datasets. The second set includes regions that are gazed by ASD instead of TD. The features of the second set will be learned by exploiting the ASD eye-tracking dataset to finetune the model. To encourage the model to precisely predict the regions gazed by most observers with ASD, we process the FDM by single-side clipping (SSC) as follows:

$$G'(i) = \begin{cases} 0, & G(i) \le T \\ G(i), & G(i) > T \end{cases} \quad (5)$$

Table 1. Comparison with state-of-the-art models on the test set with 30 images in the dataset [21]. The best two scores are marked in **bold** and underlined.

Model	SIM	CC	KL	NSS	AUC-J
GBVS	0.599	0.554	0.543	0.992	0.764
SalGAN	0.635	0.687	1.565	1.307	0.783
SAM-VGG	0.643	0.705	0.586	1.377	0.797
Our(w/o SSC)	0.671	0.734	0.465	1.459	0.808
Our	**0.678**	**0.769**	**0.421**	**1.738**	**0.834**

Table 2. Performance of our model on the evaluation set with 200 images in the dataset [21].

Model	SIM	CC	KLD	NSS	AUC-J
Our	0.623	0.681	0.590	1.510	0.818

where i indexes the i^{th} pixel, G is the original FDM that is generated from fixation map, and G' is the processed ground truth. The threshold T is set to 0.05 empirically. By using SSC, the salience of the regions where most ASD observers pay attention on are reserved, while the salience of other regions where few ASD observers gaze are depressed.

3. EXPERIMENTAL RESULTS

3.1. Experimental setup

For saliency prediction model, the backbone (i.e., VGG-16) is initialized by the pre-trained parameters on ImageNet [15]. The other parameters are initialized according to Xavier normalization [16]. First, our model was trained on the MIT1003 dataset [17] with an initial learning rate of 1e-4. The training set of ASD eye-tracking dataset [21] contains 300 images, which are split to 240, 30 and 30 for training, validation and testing, respectively. Then we finetuned the model on the first 240 images with a learning rate of 1e-6 and evaluated its performance on the validation set. Finally, the performance of our model is tested on the last 30 images. The same operations are performed for the two compared models, *i.e.* SalGAN [18] and SAM-VGG [12]. The saliency prediction maps are evaluated with five well-known metrics, *i.e.* similarity (SIM), CC, AUC Judd (AUC-J), NSS, and KL.

3.2. Comparison with state-of-the-art models

As shown in [19], by finetuning five state-of-the-art saliency prediction models on the ASD eye-tracking dataset, it was found that SalGAN [18] and SAM-VGG [12] outperform the other three models including SALICON [9], ML-Net [10] and SAM-ResNet [12]. Therefore, SalGAN and SAM-VGG are selected for comparison in this paper. Besides, the classical model GBVS [20] is also compared. As shown in Table 1, our model with SSC (the bottom line in Table 1)

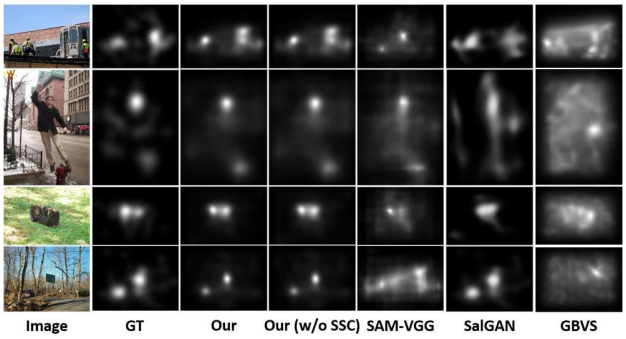

Fig. 2. Qualitative comparison with other state-of-the-art models.

outperforms all the other models. Our model without SSC still performs better than the other three models.

As shown in Table 2, the performance of our model on the evaluation set is not as good as the performance on the test set. This is explainable considering that the images in the test set are more similar to the images in the validation set than those in the evaluation set.

Some examples of saliency maps generated by all the five models are shown in Fig. 2. It can be seen that our model is capable of better capturing salient regions at different locations. Besides, our model performs better in predicting low-saliency region, and better suppresses irrelevant regions than other models.

4. CONCLUSION

In this paper, we have presented a novel saliency prediction model suitable to children with ASD. The fusion of multi-level features, deep supervision on attention maps, and the single-side clipping operation on ground truths contribute the higher prediction performance of our model Moreover, we demonstrate that it is necessary to first train the model on eye-tracking dataset of normal humans to learn salience of most attractive regions, and then finetune the model on the ASD eye-tracking dataset for more effective saliency prediction.

5. REFERENCES

[1] G. Dawson, S. J. Webb, and J. McPartland, "Understanding the nature of face processing impairment in autism: insights from behavioral and electrophysiological studies," *Developmental Neuropsychology*, vol. 27, no. 3, pp. 403-424, 2, 2005.

[2] S. Wang, M. Jiang, X. M. Duchesne, E. A. Laugeson, D. P. Kennedy, R. Adolphs, and Q. Zhao, "Atypical visual saliency in autism spectrum disorder quantified through model-based eye tracking," *Neuron*, vol. 88, no. 3, pp. 604-616, 11, 2015.

[3] N. J. Sasson, J. T. Elison, L. M. Turner-Brown, G. S. Dichter, and J. W. Bodfish, "Brief report: Circumscribed attention in young children with autism," *Journal of autism developmental disorders Autism Research*, vol. 41, no. 2, pp. 242-247, 5, 2010.

[4] N. J. Sasson, L. M. Turner-Brown, T. N. Holtzclaw, K. S. Lam, and J. W. Bodfish, " Children with autism demonstrate circumscribed attention during passive viewing of complex social and nonsocial picture arrays," *Autism Research*, vol. 1, no. 1, pp. 31-42, 2, 2008.

[5] T. Higuchi, Y. Ishizaki, A. Noritake, Y. Yanagimoto, H. Kobayashi, K. Nakamura, and K. Kaneko, "Spatiotemporal characteristics of gaze of children with autism spectrum disorders while looking at classroom scenes," *PLoS ONE*, vol. 12, no. 5, pp. 1-19, 5, 2017.

[6] T. Wadhera, and D. Kakkar, "Eye Tracker: An Assistive Tool in Diagnosis of Autism Spectrum Disorder," *Emerging Trends in the Diagnosis and Intervention of Neurodevelopmental Disorders*, pp. 125-152: IGI Global, 2018.

[7] G. Wan, X. Kong, B. Sun, S. Yu, Y. Tu, J. Park, C. Lang, M. Koh, Z. Wei, and Z. Feng, "Applying Eye Tracking to Identify Autism Spectrum Disorder in Children," *Journal of Autism Developmental Disorders*, vol. 49, no. 1, pp. 209-215, 1, 2019.

[8] M. Jiang, and Q. Zhao, "Learning visual attention to identify people with autism spectrum disorder," *in IEEE Int. Conf. on Comput. Vis.*, 2017, pp. 3267-3276.

[9] X. Huang, C. Shen, X. Boix, and Q. Zhao, "Salicon: Reducing the semantic gap in saliency prediction by adapting deep neural networks," *in IEEE Int. Conf. on Comput. Vis.*, 2015, pp. 262-270.

[10] M. Cornia, L. Baraldi, G. Serra, and R. Cucchiara, "A deep multi-level network for saliency prediction," *in Int. Conf. on Pattern Recognit.*, 2016, pp. 3488-3493.

[11] W. Wang, and J. Shen, "Deep visual attention prediction," *IEEE Transactions on Image Processing*, vol. 27, no. 5, pp. 2368-2378, May, 2018.

[12] M. Cornia, L. Baraldi, G. Serra, and R. Cucchiara, "Predicting human eye fixations via an lstm-based saliency attentive model," *IEEE Transactions on Image Processing*, vol. 27, no. 10, pp. 5142-5154, Oct, 2018.

[13] A. Borji, "Saliency prediction in the deep learning era: An empirical investigation," *arXiv preprint arXiv:1810.03716*, 10, 2018.

[14] K. Simonyan, and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 9, 2014.

[15] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, and M. Bernstein, "Imagenet large scale visual recognition challenge," *International journal of Comput. Vis.*, vol. 115, no. 3, pp. 211-252, 4, 2015.

[16] X. Glorot, and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," *in thirteenth Int. Conf. on Artificial Intelligence and Statistics*, 2010, pp. 249-256.

[17] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," *in Int. Conf. on Comput. Vis. (ICCV)*, 2009, pp. 2106-2113.

[18] J. Pan, C. C. Ferrer, K. McGuinness, N. E. OConnor, J. Torres, E. Sayrol, and X. Giro-i-Nieto, "Salgan: Visual saliency prediction with adversarial networks," *arXiv preprint arxXiv:1701.01081v3*, 01, 2017.

[19] H. Duan, G. Zhai, X. Min, Y. Fang, Z. Che, X. Yang, C. Zhi, H. Yang, and N. Liu, "Learning to Predict where the Children with Asd Look," *in Proceeding of Int. Conf. on Image Processing, 2018,* pp. 704-708.

[20] J. Harel, C. Koch, and P. Perona, "Graph-Based Visual Saliency," in *Adv. Neural Inform. Process. Syst.*, 2006, pp. 545-552.

[21] H. Duan, G. Zhai, X. Min, Z. Che, Y. Fang, X. Yang, J. Gutiérrez, and P. Le Callet, "A dataset of eye movements for the children with autism spectrum disorder," *ACM Multimedia Systems Conf.* (*MMSys '19*), Jun. 2019.

[22] F. Yu, and V. Koltun, "Multi-scale context aggregation by dilated convolutions," *arXiv preprint, arXiv: 1511.07122*, Nov. 2015.